WHAT IS CLAIMED IS:

1. A subcollection of samples from a target population, comprising:

a plurality of samples, wherein the samples are selected from the group consisting of blood, tissue, body fluid, cell, seed, microbe, pathogen and reproductive tissue samples; and

a symbology on the containers containing the samples, wherein the symbology is representative of the source and/or history of each sample, wherein:

the target population is a healthy population that has not been selected for any disease state;

the collection comprises samples from the healthy population; and

the subcollection is obtained by sorting the collection according to specified parameters.

2. The subcollection of claim 1, wherein the parameters are selected from the group consisting of ethnicity, age, gender, height, weight, alcohol intake, number of pregnancies, number of live births, vegetarians, type of physical activity, state of residence and/or length of residence in a particular state, educational level, age of parent at death, cause of parent death, former or current smoker, length of time as a smoker, frequency of smoking, occurrence of a disease in immediate family (parent, siblings, children), use of prescription drugs and/or reason therefor, length and/or number of hospital stays and exposure to environmental factors.

3. The subcollection of claim 1, wherein the symbology is a bar code.

4. A method of producing a database, comprising:

identifying healthy members of a population;

obtaining data comprising identifying information and obtaining historical information and data relating to the identified members of the population and their immediate family;

entering the data into a database for each member of the population and associating the member and the data with an indexer.

5. The method of claim 4, further comprising:

obtaining a body tissue or body fluid sample;

analyzing the body tissue or body fluid in the sample; and

entering the results of the analysis for each member into the database and associating each result with the indexer representative of each member.

6. A database produced by the method of claim 4.

7. A database produced by the method of claim 5.

8. A database, comprising:

datapoints representative of a plurality of healthy organisms from whom biological samples are obtained,

wherein each datapoint is associated with data representative of the organism type and other identifying information.

9. The database of claim 8, wherein the datapoints are answers to questions regarding one or more of a parameters selected from the group consisting of ethnicity, age, gender, height, weight, alcohol intake, number of pregnancies, number of live births, vegetarians, type of physical activity, state of residence and/or length of residence in a particular state, educational level, age of parent at death, cause of parent death, former or current smoker, length of time as a smoker, frequency of smoking, occurrence of a disease in immediate family (parent, siblings, children), use of prescription drugs and/or reason therefor, length and/or number of hospital stays and exposure to environmental factors.

10. The database of claim 9, wherein the organisms are mammals and the samples are body fluids or tissues.

11. The database of claim 9, wherein the samples are selected from blood, blood fractions, cells and subcellular organelles.

12. The database of claim 8, further comprising,

phenotypic data from an organism.

13. The database of claim 12, wherein the data includes one of physical characteristics, background data, medical data, and historical data.

14. The database of claim 8, further comprising,

genotypic data from nucleic acid obtained from an organism.

15. The database of claim 14, wherein genotypic data includes, genetic markers, non-coding regions, microsatellites, RFLPs, VNTRs, historical data of the organism, medical history, and phenotypic information.

16. The database of claim 8 that is a relational database.

17. The database of claim 16, wherein the data are related to an indexer datapoint representative of each organism from whom data is obtained.

18. A method of identifying polymorphisms that are candidate genetic markers, comprising:

identifying a polymorphism; and

identifying any pathway or gene linked to the locus of the polymorphism, wherein

the polymorphisms are identified in samples associated with a target population that comprises healthy subjects.

19. The method of claim 18, wherein the polymorphism is identified by detecting the presence of target nucleic acids in a sample by a method, comprising the steps of:

a) hybridizing a first oligonucleotide to the target nucleic acid;

b) hybridizing a second oligonucleotide to an adjacent region of the target nucleic acid;

c) ligating the hybridized oligonucleotides; and

c) detecting hybridized first oligonucleotide by mass spectrometry as an indication of the presence of the target nucleic acid.

20. The method of claim 18, wherein the polymorphism is identified by detecting target nucleic acids in a sample by a method, comprising the steps of:

a) hybridizing a first oligonucleotide to the target nucleic acid and hybridizing a second oligonucleotide to an adjacent region of the target nucleic acid;

b) contacting the hybridized first and second oligonucleotides with a cleavage enzyme to form a cleavage product; and

c) detecting the cleavage product by mass spectrometry as an indication of the presence of the target nucleic acid.

21. The method of claim 20 wherein the samples are from subjects in a healthy database.

22. The method of claim 18, wherein the polymorphism is identified by identifying target nucleic acids in a sample by primer oligo base extension (probe).

23. The method of 22, wherein primer oligo base extension, comprises:

a) obtaining a nucleic acid molecule that contains a target nucleotide;

b) optionally immobilizing the nucleic acid molecule onto a solid support, to produce an immobilized nucleic acid molecule;

c) hybridizing the nucleic acid molecule with a primer oligonucleotide that is complementary to the nucleic acid molecule at a site adjacent to the target nucleotide;

d) contacting the product of step c) with a composition comprising a dideoxynucleoside triphosphate or a 3'-deoxynucleoside triphosphates and a polymerase, so that only a dideoxynucleoside or 3'-deoxynucleoside triphosphate that is complementary to the target nucleotide is extended onto the primer; and

e) detecting the extended primer, thereby identifying the target nucleotide.

24. The method of claim 23, wherein detection of the extended primer is effected by mass spectrometry, comprising:

ionizing and volatizing the product of step d) ; and

detecting the extended primer by mass spectrometry, thereby identifying the target nucleotide.

25. The method of claim 24, wherein;

samples are presented to the mass spectrometer as arrays on chips; and

each sample occupies a volume that is about the size of the laser spot projected by the laser in a mass spectrometer used in matrix-assisted laser desorption/ionization (MALDI) spectrometry.

26. A combination, comprising:

a database containing parameters associated with a datapoint representative of a subject from whom samples are obtained, wherein the subjects are healthy; and

an indexed collection of the samples, wherein the index identifies the subject from whom the sample was obtained.

27 The combination of claim 26, wherein the parameter is selected from the group consisting of ethnicity, age, gender, height, weight, alcohol intake, number of pregnancies, number of live births, vegetarians, type of physical activity, state of residence and/or length of residence in a particular state, educational level, age of parent at death, cause of parent death, former or current smoker, length of time as a smoker, frequency of smoking, occurrence of disease in immediate family (parent, siblings, children), use of prescription drugs and/or reason therefor, length and/or number of hospital stays and ecposure to environmental factors.

28. The combination of claim 26, wherein the database further contains genotypic data for each subject.

29. The combination of claim 26, wherein the samples are blood.

30 A data storage medium, comprising the database of claim 8.

31. A computer system, comprising the database of claim 8.

32. A system for high throughput processing of biological samples, comprising:

a process line comprising a plurality of processing stations, each of which performs a procedure on a biological sample contained in a reaction vessel;

a robotic system that transports the reaction vessel from processing station to processing station;

a data analysis system that receives test results of the process line and automatically processes the test results to make a determination regarding the biological sample in the reaction vessel;

a control system that determines when the test at each processing station is complete and, in response, moves the reaction vessel to

the next test station, and continuously processes reaction vessels one after another until the control system receives a stop instruction; and

a database of claim 8, wherein the samples tested by the automated process line comprise samples from subjects in the database.

33. The system of claim 32, wherein one of the processing stations comprises a mass spectrometer.

34. The system of claim 32, wherein the data analysis system processes the test results by receiving test data from the mass spectrometer such that the test data for a biological sample contains one or more signals, whereupon the data analysis system determines the area under the curve of each signal and normalizes the results thereof and obtains a substantially quantitative result representative of the relative amounts of components in the tested sample.

35. A method for high throughput processing of biological samples, the method comprising:

transporting a reaction vessel along a system of claim 32, comprising a process line having a plurality of processing stations, each of which performs a procedure on one or more biological samples contained in the reaction vessel;

determining when the test procedure at each processing station is complete and, in response, moving the reaction vessel to the next processing station;

receiving test results of the process line and automatically processing the test results to make a data analysis determination regarding the biological samples in the reaction vessel; and

processing reaction vessels continuously one after another until receiving a stop instruction, wherein the samples tested by the automated process line comprise samples from subjects in the database.

36. The method of 35, wherein one of the processing stations comprises a mass spectrometer.

37. The method of claim 36, wherein the samples are analyzed by a method comprising primer oligo base extension (probe).

38. The method of claim 37, further comprising:

processing the test results by receiving test data from the mass spectrometer such that the test data for a biological sample contains one or more signals or numerical values representative of signals, whereupon the data analysis system determines the area under the curve of each signal and normalizes the results thereof and obtains a substantially quantitative result representative of the relative amounts of components in the tested sample.

39. The method of claim 37, wherein primer oligo base extension, comprises:

a) obtaining a nucleic acid molecule that contains a target nucleotide;

b) optionally immobilizing the nucleic acid molecule onto a solid support, to produce an immobilized nucleic acid molecule;

c) hybridizing the nucleic acid molecule with a primer oligonucleotide that is complementary to the nucleic acid molecule at a site adjacent to the target nucleotide;

d) contacting the product of step c) with composition comprising a dideoxynucleoside triphosphate or a 3'-deoxynucleoside triphosphates and a polymerase, so that only a dideoxynucleoside or 3'-deoxynucleoside triphosphate that is complementary to the target nucleotide is extended onto the primer; and

e) detecting the primer, thereby identifying the target nucleotide.

40. The method of 39, wherein detection of the extended primer is effected by mass spectrometry, comprising:

ionizing and volatizing the product of step d); and

detecting the extended primer by mass spectrometry, thereby identifying the target nucleotide.

41. The method of claim 36, wherein the target nucleic acids in the sample are detected and/or identified by a method, comprising the steps of:

a) hybridizing a first oligonucleotide to the target nucleic acid;

b) hybridizing a second oligonucleotide to an adjacent region of the target nucleic acid;

c) ligating then hybridized oligonucleotides; and

c) detecting hybridized first oligonucleotide by mass spectrometry as an indication of the presence of the target nucleic acid.

42. The method of claim 36, wherein the target nucleic acids in the sample are detected and/or identified by a method, comprising the steps of:

a) hybridizing a first oligonucleotide to the target nucleic acid and hybridizing a second oligonucleotide to an adjacent region of the target nucleic acid;

b) contacting the hybridized first and second oligonucleotides with a cleavage enzyme to form a cleavage product; and

c) detecting the cleavage product by mass spectrometry as an indication of the presence of the target nucleic acid.

SUB D4

43. A method of producing a database stored in a computer memory, comprising:

identifying healthy members of a population;

obtaining identifying and historical information and data relating to the identified members of the population;

entering the member-related data into the computer memory database for each identified member of the population and associating the member and the data with an indexer.

44. The method of claim 43, further comprising:

obtaining a body tissue or body fluid sample of an identified member;

analyzing the body tissue or body fluid in the sample; and

entering the results of the analysis for each member into the computer memory database and associating each result with the indexer representative of each member.

45. A database produced by the method of claim 43.

46. A database produced by the method of claim 44.

47. The database of claim 8, wherein:

the organims are selected from among animals, bacteria, fungi, protozoans and parasites and

each datapoint is associated with parameters representative of the organism type and identifying information.

48. The database of claim 43, further comprising, phenotypic data regarding each subject.

49. The database of claim 47 that is a relational database and the parameters are the answers to the questions in the questionnaire.

50. The database of claim 8, further comprising,

genotypic data of nucleic acid of the subject, wherein genotypic data includes, but is not limited to, genetic markers, non-coding regions, microsatellites, restriction fragment length polymorphisms (RFLPs), variable number tandem repeats (VNTRs), historical day of the organism, the medical history of the subject, phenotypic information, and other information.

51. A database, comprising data records stored in computer memory, wherein the data records contain information that identifies healthy members of a population, and also contain identifying and historical information and data relating to the identified members.

52. The database of claim 51, further comprising an index value for each identified member that associates each member of the population with the identifying and historical information and data.

53. A computer system, comprising the database of claim 51.

54. An automated process line, comprising the database of claim 51.

55. A method for determining a polymorphism that correlates with age, ethnicity or gender, comprising:

identifying a polymorphism; and

determining the frequency of the polymorphism with increasing age, with ethnicity or with gender in a healthy population.

56. A method for determining whether a polymorphism correlates with suceptibility to morbidity, early mortality, or morbidity and early mortality, comprising;

identifying a polymorphism; and

determining the frequency of the polymorphism with increasing age in a healthy population.

57. A high throughput method of determining frequencies of genetic variations, comprising:

selecting a healthy target population and a genetic variation to be assessed;

pooling a plurality of samples of biopolymers obtained from members of the population,

determining or detecting the biopolymer that comprises the variation by mass spectrometry;

obtaining a mass spectrum or a digital representation thereof; and

determining the frequency of the variation in the population.

58. The method of claim 57, wherein:

the variation is selected from the group consisting of an allelic variation, a post-translational modification, a nucleic modification, a label, a mass modification of a nucleic acid and methylation; and/or

the biopolymer is a nucleic acid, a protein, a polysaccharide, a lipid, a small organic metabolite or intermediate, wherein the concentration of biopolymer of interest is the same in each of the samples; and/or

the frequency is determined by assessing the method comprising determining the area under the peak in the mass spectrum or digital repesentation thereof corresponding to the mass of the biopolymer comprising the genomic variation.

59. The method of claim 58, wherein the method for determining the frequency is effected by determining the ratio of the signal or the digital representation thereof to the total area of the entire mass spectrum, which is corrected for background.

60. A method for discovery of a polymorphism in a population, comprising:

sorting the database of claim 8 according to a selected parameter to identify samples that match the selected parameter;

isolating a biopolymer from each identified sample;

optionally pooling each isolated biopolymer;

optionally amplifying the amount of biopolymer;

cleaving the pooled biopolymers to produce fragments thereof;

obtaining a mass spectrum of the resulting fragments and comparing the mass spectrum with a control mass spectrum to identify differences between the spectra and thereby identifing any polymorphisms; wherein:

the control mass spectrum is obtained from unsorted samples in the collection or samples sorted according to a different parameter.

61. The method of claim 60, wherein cleaving is effected by contacting the biopolymer with an enzyme.

62. The method of claim 61, wherein the enzyme is selected from the group consisting of nucleotide glycosylase, a nickase and a type IIS restriction enzyme.

63. The method of claim 60, wherein the biopolymer is a nucleic acid or a protein.

64. The method of claim 60, wherein the the mass spectrometric format is selected from among Matrix-Assisted Laser Desorption/Ionization, Time-of-Flight (MALDI-TOF), Electrospray (ES), IR-MALDI, Ion Cyclotron Resonance (ICR), Fourier Transform and combinations thereof.

65. A method for discovery of a polymorphism in a population, comprising:

obtaining samples of body tissue or fluid from a plurality of organisms;

isolating a biopolymer from each sample;

pooling each isolated biopolymer;

optionally amplifying the amount of biopolymer;

cleaving the pooled biopolymers to produce fragments thereof;

obtaining a mass spectrum of the resulting fragments;

comparing the frequency of each fragment to identify fragments present in amounts lower than the average frequency, thereby identifying any polymorphisms.

66. The method of claim 65, wherein cleaving is effected by contacting the biopolymer with an enzyme.

67. The method of claim 66, wherein the enzyme is selected from the group consisting of nucleotide glycosylase, a nickase and a type IIS restriction enzyme.

68. The method of claim 65, wherein the biopolymer is a nucleic acid or a protein.

69. The method of claim 65, wherein the the mass spectrometric format is selected from among Matrix-Assisted Laser Desorption/Ionization, Time-of-Flight (MALDI-TOF), Electrospray (ES), IR-MALDI, Ion Cyclotron Resonance (ICR), Fourier Transform and combinations thereof.

70. The method of claim 65, wherein the samples are obtained from healthy subjects.

71. A method of correlating a polymorphism with a parameter, comprising:

sorting the database of claim 8 according to a selected parameter to identify samples that match the selected parameter;

isolating a biopolymer from each identified sample;

pooling each isolated biopolymer;

optionally amplifying the amount of biopolymer;

determining the frequency of the polymorphism in the pooled biopolymers, wherein:

an alteration of the frequency of the polymorphism compared to a control, indicates a correlation of the polymorphism with the selected parameter; and

the control is the frequency of the polymorphism in pooled biopolymers obtained from samples identified from an unsorted database or from a database sorting according to a different parameter.

72. The method claim 71, wherein the parameter is selected from the group consisting of ethnicity, age, gender, height, weight, alcohol intake, number of pregnancies, number of live births, vegetarians, type of physical activity, state of residence and/or length of residence in a particular state, educational level, age of parent at death, cause of parent death, former or current smoker, length of time as a smoker, frequency of smoking, occurrence of a disease in immediate family (parent, siblings, children), use of prescription

drugs and/or reason therefor, length and/or number of hospital stays and exposure to environmental factors.

73. The method claim 72, wherein the parameter is occurrence of disease or a particular disease in an immediate family member, thereby correlating the polymorphism with the disease.

74. The method of claim 71, wherein the pooled biopolymers are pooled nucleic acid molecules.

75. The method of claim 74, wherein the polymorphism is detected by primer oligo base extension (PROBE).

76. The method of 75, wherein primer oligo base extension, comprises:

a) optionally immobilizing the nucleic acid molecules onto a solid support, to produce immobilized nucleic acid molecules;

b) hybridizing the nucleic acid molecules with a primer oligonucleotide that is complementary to the nucleic acid molecule at a site adjacent to the polymorphism;

c) contacting the product of step c) with composition comprising a dideoxynucleoside triphosphate or a 3'-deoxynucleoside triphosphates and a polymerase, so that only a dideoxynucleoside or 3'-deoxynucleoside triphosphate that is complementary to the polymorphism is extended onto the primer; and

d) detecting the extended primer, thereby detecting the polymorphism in nucleic acid molecules in the pooled nucleic acids.

77. The method of claim 76, wherein detecting is effected by mass spectrometry.

78. The method of claim 71, wherein the frequency is percentage of nucleic acid molecules in the pooled nucleic acids that contain the polymorphism.

79. The method of claim 78, wherein the ratio is determined by obtaining mass spectra of the pooled nucleic acids.

80. The method of claim 72, wherein the parameter is age, thereby correlating the polymorphism with suceptibility to morbidity, early mortality or morbidity and early mortality.

81. A method for haplotyping polymorphisms in a nucleic acid, comprising:

(a) sorting the database of claim 8 according to a selected parameter to identify samples that match the selected parameter;

(b) isolating nucleic acid from each identified sample;

(c) optionally pooling each isolated nucleic acid;

(d) amplifying the amount of nucleic acid;

(e) forming single-stranded nucleic acid and splitting each single-strand into a separate reaction vessel;

(f) contacting each single-stranded nucleic acid with an adaptor nucleic acid to form an adaptor complex;

(g) contacting the adaptor complex with a nuclease and a ligase;

(h) contacting the products of step (g) with a mixture that is capable of amplifying a ligated adaptor to produce an extended product;

(i) obtaining a mass spectrum of each nucleic acid resulting from step (h) and detecting a polymorphism by identifying a signal corresponding to the extended product;

(j) repeating steps (f) through (i) utilizing an adaptor nucleic acid able to hybridize with another adapter nucleic acid that hybridizes to a different sequence on the same strand; whereby

the polymorphisms are haplotyped by detecting more than one extended product.

82. The method of claim 1, wherein the nuclease is Fen-1.

83. A method for haplotyping polymorphisms in a population, comprising:

sorting the database of claim 8 according to a selected parameter to identify samples that match the selected parameter;

isolating a nucleic acid from each identified sample;

pooling each isolated nucleic acid;

optionally amplifying the amount of nucleic acid;

contacting the nucleic acid with at least one enzyme to produce fragments thereof;

obtaining a mass spectrum of the resulting fragments; whereby:

the polymorphisms are detected by detecting signals corresponding to the polymorphisms; and

the polymorphisms are haplotyped by determining from the mass spectrum that the polymorphisms are located on the same strand of the nucleic acid.

84. The method of claim 83, wherein the enzyme is a nickase.

85. The method of claim 84, wherein the nickase is selected from the group consisting of NY2A and NYS1.

86. A method for detecting methylated nucleotides within a nucleic acid sample, comprising:

splitting a nucleic acid sample into separate reaction vessels;

contacting nucleic acid in one reaction vessel with bisulfite;

amplifying the nucleic acid in each reaction vessel;

cleaving the nucleic acids in each reaction vessel to produce fragments thereof;

obtaining a mass spectrum of the resulting fragments from one reaction vessel and another mass spectrum of the resulting fragements from another reaction vessel; whereby:

cytosine methylation is detected by identifying a difference in signals between the mass spectra.

87. The method of claim 86, wherein:

the step of amplifying is carried out in the presence of uracil; and

the step of cleaving is effected by a uracil glycosylase.

88. A method for identifying a biological sample, comprising:

generating a data set indicative of the composition of the biological sample;

denoising the data set to generate denoised data;

deleting the baseline from the denoised data to generate an intermediate data set;

defining putative peaks for the biological sample;

using the putative peaks to generate a residual baseline;

removing the residual baseline from the intermediate data set to generate a corrected data set;

locating, responsive to removing the residual baseline, a probable peak in the corrected data set; and

identifying, using the located probable peak, the biological sample;

wherein the generated biological sample data set comprises data from sense strands and antisense strands of assay fragments.

89. The method according to claim 88, wherein identifying includes combining data from the sense strands and the antisense strands, and comparing the data against expected sense strand and antisense strand values, to identify the biological sample.

90. The method according to claim 88, wherein identifying includes deriving a peak probability for the probable peak, in accordance with whether the probable peak is from sense strand data or from antisense strand data.

91. The method according to claim 88, wherein identifying includes deriving a peak probability for the probable peak and applying an allelic penalty in response to a ratio between a calculated area under the probable peak and a calculated expected average area under all peaks in the data set.

92. A method for identifying a biological sample, comprising:

generating a data set indicative of the composition of the biological sample;

denoising the data set to generate denoised data;

deleting the baseline from the denoised data to generate an intermediate data set;

defining putative peaks for the biological sample; using the putative peaks to generate a residual baseline;

removing the residual baseline from the intermediate data set to generate a corrected data set;

locating, responsive to removing the residual baseline, a probable peak in the corrected data set; and

identifying, using the located probable peak, the biological sample;

wherein identifying includes deriving a peak probability for the probable peak and

applying an allelic penalty in response to a ratio between a calculated area under the probable peak and a calculated expected average area under all peaks in the data set.

93. The method according to claim 92, wherein identifying includes comparing data from probable peaks that did not receive an applied allelic penalty to determine their mass in accordance with oligonucleotide biological data.

94. The method according to claim 92, wherein the allelic penalty is not applied to probable peaks whose ratio of area under the peak to the expected area value is greater than 30%.

95. A method for detecting a polymorphism in a nucleic acid, comprising:

amplifying a region of the nucleic acid to produce an amplicon, wherein the resulting amplicon comprises one or more enzyme restriction sites;

contacting the amplicon with a restriction enzyme to produce fragments;

obtaining a mass spectrum of the resulting fragments and analyzing signals in the mass spectrum by the method of claim 88; whereby:

the polymorphism is detected from the pattern of the signals.

96. A subcollection of samples from a target population, comprising:

a plurality of samples, wherein the samples are selected from the group consisting of nucleic acids, fetal tissue, protein samples; and

a symbology on the containers containing the samples, wherein the symbology is representative of the source and/or history of each sample, wherein:

the target population is a healthy population that has not been selected for any disease state;

the collection comprises samples from the healthy population; and

the subcollection is obtained by sorting the collection according to specified parameters.

97. The combination of claim 26, wherein the samples are selected selected from the group consisting of nucleic acids, fetal tissue, protein, tissue, body fluid, cell, seed, microbe, pathogen and reproductive tissue samples.

98. A combination, comprising the database of claim 8 and a mass spectrometer.

99. The combination of claim 98 that is an automated process line for analyzing biological samples.

100. A system for high throughput processing of biological samples, comprising:

a database of claim 8, wherein the samples tested by the automated process line comprise samples from subjects in the database; and

a mass spectrometry for analysis of biopolymers in the samples.